

A Rule-Based Speech Morphing for Verifying a Expressive Speech Perception Model

Chun-Fang Huang and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
 chuang@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

This paper describes a rule-based approach for verifying a three-layer model that was proposed for modeling expressive speech perception. The three layers are expressive speech, semantic primitives, and acoustic features. In our previous work we built the model. In the current work, the built model is verified by creating rules with parameters that morph the acoustic characteristics of a neutral utterance to the perception of certain semantic primitives or expressive speech categories. There are two types of rules. Base rules verify the validity of the analytic results. Intensity rules verify the perceived intensity of expressive speech and semantic primitives. The experiments results show the significant relationships of expressive speech, semantic primitives, and acoustic features. This model will help to develop tools such as a synthesizer to produce utterances that could give listeners the perception of different categories and intensity-levels of expressive speech.

Index Terms: expressive speech, perception, multi-layer model, fuzzy inference system, acoustic analysis

1. Introduction

Previous work on expressive speech focused on two major research areas: recognition and synthesis. Expressive speech recognition is a method for automatically identifying the emotional state of speakers [1]. Expressive speech synthesis involves manipulating parameters of the acoustic features of the voice data to produce perception of different emotions by receivers [2]. Work heretofore has been concentrated on measuring acoustic features in the speech signal and then using statistical methodology to select the most significant features for the classification or identification of emotions.

However, results may be limited by this method due to certain factors: (1) only limited types of acoustic features are studied, (2) the resulting relationships are largely dependent on the particular voice/speaker studied, (3) reproduction of expressive speech either in terms of recognition or synthesis, has not been effectively achieved, even though it is known that certain categories of expressive speech may be characterized by specific acoustic features [3]. A model to approach the study of expressive speech is needed [2]

The purpose of this study is to propose an effective model that could explain expressive speech perception. The concept of the model is based on two assumptions (1) an emotion a human being perceives in speech may depend on the "semantic primitives" sensed, such as bright or high, and (2) the perception of emotion involves humans' awareness of many vague perceptions. To explore the ideas, our previous work [4] proposed a three-layer model as shown in Figure 1. Five categories of expressive speech, Neutral, Joy, Sadness, Cold Anger, and Hot Anger, constitute the top layer, semantic

primitives constitute the middle layer, and acoustic features the bottom layer.

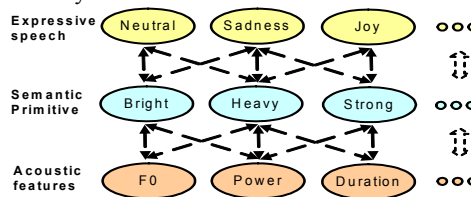


Figure 1: The conceptual diagram of the model

The model is constructed by two stages. The first stage is to build the model by analysis and the second stage is to evaluate the model by resynthesis. In this paper we describe the work of the second stage. The first stage is briefly reviewed here; the details can be found in [4]

In the first stage, two relationships of the model were built. First, the relationship between expressive speech and semantic primitives was built by conducting experiments to investigate speech in terms of the expressive speech categories, and then selecting suitable semantic primitives. According to experiment results and MDS analysis, a total of 17 semantic primitives were selected (see Table 1), which were then applied to a fuzzy inference system to model the relationship between expressive speech categories and semantic primitives. Second, the relationship between semantic primitive and acoustic feature was built by analyzing the acoustic features of the speech signal, and then building the relationship according to correlation analysis. A total of 16 acoustic features were measured: Four involved F0--mean value of rising slope (RS), highest pitch (HP), average pitch (AP) and rising slope of the first accentual phrase (RS1st); four involved power envelope--mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (PRS1st), the ratio between the average power in high frequency portion (over 3 kHz) and the average power (RHT); five involved the power spectrum--first formant frequency (F1), second formant frequency (F2), third formant frequency (F3), spectral tilt (SPTL), spectral balance (SB); and three involved duration total length (TL), consonant length (CL), ratio between consonant length and vowel length (RCV). Combining the two relationships, a perceptual model as shown in Figure 2 was built for each expressive speech category.

In the second stage, as reported in this paper, we verify the built relationships. The verification method is to create rules with parameters that morph the acoustic characteristics of a neutral utterance to the perception of certain semantic primitives or expressive speech categories. These parameters come from analytic results of these two relationships. Different from other research studies, we examine the various types of human perception of emotion as well as the intensity

of the perceptions from the resulting models. That is, there will be a rule for morphing a neutral utterance to a semantic primitive and from there, to an expressive speech category. Then the parameters will be modified to create variations of the rules that should give different intensity-levels of perception.

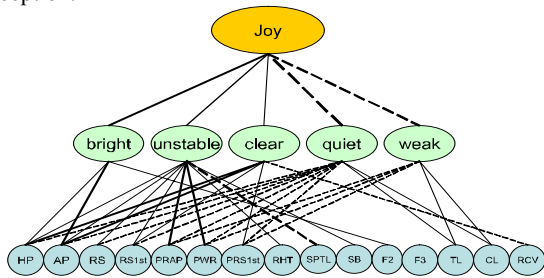


Figure 2: Resultant perceptual model of Joy. The solid lines indicate the relation is a positive correlation, and the dotted ones indicate a negative correlation. The thicker the line is, the higher the correlation.

The three-layered model outlined in this research is a new and different point of view for expressive speech perception. New in this approach is the concept of semantic primitives, used as the middle layer in the model. The morphing rules created from the resulting models not only giving listeners the perception of different semantic primitives and expressive speech categories but also their different intensity levels.

2. Verification of the perceptual model

2.1. The Verification of the relationship between semantic primitives and acoustic features

Two types of semantic-primitive rules (i.e., morphing rules) were developed. (1) The base rules morphed a neutral utterance into an utterance which could be perceived in terms of one and only one semantic primitive. In this way it was possible to assess which acoustic features are involved in creating the percept of each semantic primitive. (2) The intensity rules involved morphing rules which morphed an utterance in such a way that the intensity of the semantic primitive changed. In this way, it was possible to assess how a change in the intensity of the acoustic features changed the intensity levels of semantic primitives.

2.1.1. Based rule development

Base rules, developed from the analysis of acoustic features, are such that there is one base rule for one semantic primitive. One rule has 16 parameters which control the 16 acoustic features. To decide the parameters of the base rules, 9 utterances for each semantic primitive were selected according to the results of previous experiments [4] that were well-perceived for that semantic primitive. The parameters of rules were decided by calculating the percentage variation of the acoustic features between the selected utterances and their corresponding neutral utterances. The equation is

$$\frac{vaf_i - vnaf_i}{vnaf_i} \quad (1)$$

where vaf_i is the value of acoustic features of i th utterance and $vnaf_i$ is the value of the corresponding neutral utterance.

2.1.2. Intensity rule development

The intensity rules were created by adjusting the parameters of the base rules. According to the resulting models as shown in Figure 2, we created three intensity rules (SR1, SR2, and SR3). SR1 is the base rule. The utterance morphed by SR2 was supposed to be with stronger perception than that morphed by SR1; the utterance morphed by SR3 was supposed to be with stronger perception than that morphed by SR2. Specifically, SR2 was created by increasing parameters of SR1 with 4% or 2% for the solid thick and thin lines, respectively, or by decreasing with 4% or 2% for the dotted thick and dotted thin lines, respectively. SR3 was created by increasing parameters of SR2 with 4% or 2% for the solid thick and thin lines, respectively or by decreasing with 4% or 2% for the dotted thick and dotted thin lines, respectively. Thus, the solid lines indicate a positive correlation, and the dotted ones, a negative correlation. The thicker the line is, the higher the correlation. In this way, the parameters of the base rules were adjusted such that the parameter with a solid thick line would be changed in a positive direction by a larger amount than that of the solid thin line. The parameter with a dotted thick line would be changed in a negative direction by a larger amount than the dotted thin line.

2.1.3. Rule Implementation

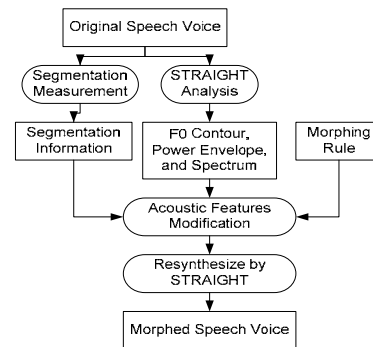


Figure 3: Process of morphing voices in STRAIGHT

A morphing process was developed (see Figure 3) for rule implementation. F0 contour, power envelope, and spectrum were extracted from the neutral speech signal by using STRAIGHT while segmentation information was measured manually. Next, acoustic features in terms of F0 contour, power envelope, spectrum and duration were modified according to the rules. Finally, we used STRAIGHT to re-synthesize the expressive speech utterance using the modified F0 contour, power envelope, spectrum [5] and duration.

2.1.4. Experiments and discussion

Two experiments are conducted to examine the base rules and the intensity rules, respectively. In the first experiment, 17 morphed speech utterances were produced by implementing the created semantic primitive base rules, giving one morphed speech utterance for each semantic primitive. In addition, there was one neutral speech utterance. Subjects were ten Japanese graduate students with normal hearing ability. Subjects were asked to compare (a) a morphed speech utterance with (b) the neutral speech utterance and to choose which utterance was most associated with a particular semantic primitive. The question was “Is (a) or (b) more ‘bright’?” Paired stimuli were presented randomly to each subject through binaural headphones at a comfortable sound pressure level. The accuracy rate results shown in Table 1

indicate that most of the morphed speech utterances were perceived as the semantic-primitive intended by the morphed speech utterance. The semantic primitive Monotonous showed the lowest rate of accuracy. Perhaps this was because Monotonous is most similar to the neutral voice, and it is difficult to morph a monotonous utterance into a “more” monotonous utterance. These results suggest that the created base rules are effective.

Table 1. *Experimental results of base rule evaluation.*

Semantic Primitive	Accuracy Rates	Semantic Primitive	Accuracy Rates
bright	100%	monotonous	60%
dark	100%	heavy	90%
high	100%	clear	80%
low	100%	noisy	100%
strong	100%	quiet	90%
weak	80%	sharp	90%
calm	100%	fast	100%
unstable	100%	slow	100%
well-modulated	100%		

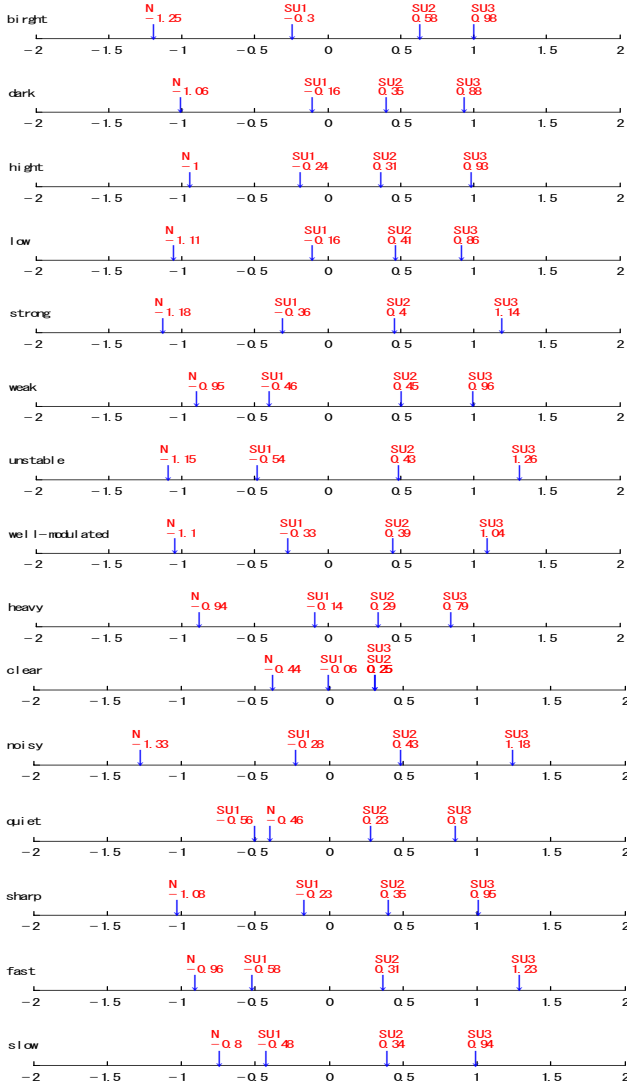


Figure 4: *Experimental results of intensity rule evaluation. The intended intensity levels are SU3 > SU2 > SU1 > N. That is, SU3 should have a higher level of intensity*

perception than SU2 than SU1 than N. N is the neutral utterance.

In the second experiment, stimuli were three morphed utterances (SU1, SU2, and SU3) for each semantic primitive and one neutral utterance. SU1, SU2, and SU3 were morphed by the intensity rules SR1, SR2, and SR3, respectively. Scheffe’s method of paired comparison was used to evaluate the intensity of the semantic primitive. Subjects were asked to evaluate which stimulus had a stronger intensity of the semantic primitive according to a five-grade scale.

Figure 4 shows the results of the experiment. The numbers under the horizontal axis indicate the intensity levels of the semantic-primitive. The results show that listeners were able to perceive four levels of intensity for each semantic primitive, except for quiet, for which only three levels were perceived. The difficulty in perceiving different intensity levels for quiet could be because the neutral utterances are intrinsically quiet. These results suggest that by adjusting the parameters of the semantic-primitive rules, it is possible to control the intensity of the perception of the semantic primitives. They also suggest that the relationship between semantic primitives and acoustic features is a valid one.

2.2. The verification of the relationship between expressive speech and semantic primitives

In the same manner as described above, verification of the relationship between expressive speech and semantic primitives also involved rule-based speech morphing and perceptual experiments. Two types of expressive speech rules were also developed: (1) The base rules involved rules to morph a neutral utterance into an utterance which could be perceived in terms of one expressive speech category, and (2) the intensity rules involved rules to morph an utterance in such a way that the intensity of the expressive speech category changed.

2.2.1. Base rule development

The base rules of the semantic primitives were combined to form base rules for each expressive speech category. How the rules were combined is determined from the relationship between expressive speech and semantic primitives. As shown in Figure 2, the widths of the lines between the two layers of the model represent the weight values of the combinations. The weight value is higher for a thicker line and lower for a thinner line. The values of these weight combinations are shown in Table 2. For example, the base rule for Joy is calculated by adding the various base rules of the appropriate semantic primitives, and then multiplying these by the appropriate weight values as shown below: $Joy = (Bright * 0.101 + Unstable * 0.063 + Clear * 0.034 + Quiet * (-0.039) + Weak * (-0.036)) / 0.123$

Table 2. *The relationship between expressive speech and semantic primitive represented in weight values [4].*

SP	Neutral		Joy		Cold Anger		Sadness		Hot Anger		
	W	SP	W	SP	W	SP	W	SP	W	SP	
monotonous	0.270	bright	0.101	heavy	0.197	heavy	0.074	well-modulated	0.124		
clear	0.127	unstable	0.063	well-modulated	0.091	weak	0.065	unstable	0.120		
calm	0.103	clear	0.034	low	0.090	quiet	0.057	sharp	0.103		
heavy	-0.329	quiet	-0.039	slow	-0.231	strong	-0.049	calm	-0.063		
weak	-0.181	weak	-0.036	clear	-0.062	sharp	-0.079	quiet	-0.047		

2.2.2. Intensity rule development

Three intensity rules for each expressive speech category (ER1, ER2, and ER3) were created, where ER2 would generate a speech utterance that was perceived more strongly than that generated by ER1, and ER3, more than ER2. Note that for the category Neutral, there were no intensity rules. These intensity rules were created by combining intensity rules of semantic primitives. Table 3 shows an example of a way of combining intensity rules. As can be seen from Figure 2, Joy is positively correlated with Bright, Unstable and Clear, but negatively correlated with Heavy and Weak. Therefore, ER1 for Joy can be created by combining the intensity rule SR1 of Bright, SR1 of Clear, SR1 of Unstable, SR3 of Heavy, and SR3 of Weak. Along similar lines, ER2 for Joy can be created by combining the intensity rules SR2, SR2, SR2, SR2, and SR2 of Bright, Clear, Calm, Heavy, and Weak, respectively. Other intensive rules are also created in the same manner. We use the same weight combination values when creating semantic-primitive base rules for combining the semantic-primitive intensity rules here.

Table 3. An example of semantic primitive rule combination for Joy.

Joy	Bright	Unstable	Clear	Heavy	Weak
ER1	SR1	SR1	SR1	SR3	SR3
ER2	SR2	SR2	SR2	SR2	SR2
ER3	SR3	SR3	SR3	SR1	SR1

2.2.3. Experiments and discussion

Two experiments were conducted to examine the base rules and intensity rules, respectively. The experimental conditions including subjects were the same as in the previous experiments. The stimuli were one neutral utterance and 4 morphed utterances that were morphed from the neutral utterance, one for each expressive speech category. Subjects were asked to compare (a) a morphed utterance with (b) the neutral voice and to choose which utterance was most associated with a specific expressive speech category. The questions asked were like “Is (a) or (b) more ‘joyful’?”

Table 4 shows that the morphed speech utterances were perceived as the expressive speech category intended by the morphing process--100% accuracy rate for each of the expressive speech categories, except Sad, (90%). This result suggests that the created base rules are effective, and moreover, that the combinations are appropriate.

Table 4. Experiment results of base rule evaluation

Expressive Speech Category	Accuracy Rate
Joy	100%
Cold Anger	100%
Sadness	90%
Hot Anger	100%

In the second experiment, stimuli include three morphed utterances for each expressive speech category plus one neutral utterance, where the three morphed utterances are with three different levels of intensity perception (EU1, EU2, and EU3). EU2 should have stronger perception than EU1, and EU3 should have stronger perception than EU2. Scheffe’s method of paired comparison was used to evaluate the semantic primitive intensity of the utterances. Subjects were the same as in the previous experiments, and evaluated the intensity of each utterance according to a five-scale rating.

The results in Figure 5 show the four stimuli of each expressive speech category listed in ascending order of the perception of intensity. The order is congruent with what was intended by the intensity rules. These results suggest that it is possible to control the intensity of emotional perception by adjusting the intensity of semantic primitives. Moreover, the perception of expressive speech categories appears to be related to the perception of semantic primitives. These results lend validity to the model we propose here in that it substantiates the relationship between semantic primitives and expressive speech categories.

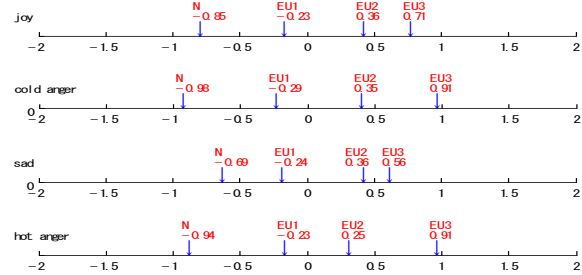


Figure 5: Experiment result of intensity rule evaluation

3. Conclusions

This paper reports a method of rule-based speech morphing for verifying an expressive speech perception model we have proposed before. We use two types of rules, base rules and intensity rules, to verify the two relationships of the model. The results show that by using the model we built it is possible to create morphing rules to control the perception of certain semantic primitives and expressive speech categories, and even their intensity levels.

This model will help develop better tools for expressive speech synthesis and recognition, for instance, a synthesizer to produce utterances that could give listeners the perception of different categories and intensity-levels of expressive speech by only controlling the combinations of various semantic primitives.

4. Acknowledgements

This research is conducted as a program for the “21st Century COE Program” by Ministry of Education, Culture, Sports, Science and Technology. This study was also supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan.

5. References

- [1] Ververidis, D and Kotropoulos, C., "Expressive speech recognition: resources, features, and methods", Speech Communication 48, 2006.
- [2] Scherer, K. R., “Vocal communication of emotion: a review of research paradigms”, Speech Communication 40, 2003, 227-256.
- [3] Sobol Shikler T. and Robinson P., “Visualizing dynamic features of expressions in speech”, In Proc. ICSLP2004, Korea.
- [4] Huang, C. F and Akagi, M., “A Multi-Layer Fuzzy Logical Model for Emotional Speech Perception”, In Proc. Interspeech 2004, Portugal.
- [5] Nguyen B. H. and Akagi, M., “A Flexible Spectral Modification Method based on Temporal Decomposition and Gaussian Mixture Model,” submitted to InterSpeech, 2007.